PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-175305
(43)Date of publication of application : 21.06.2002

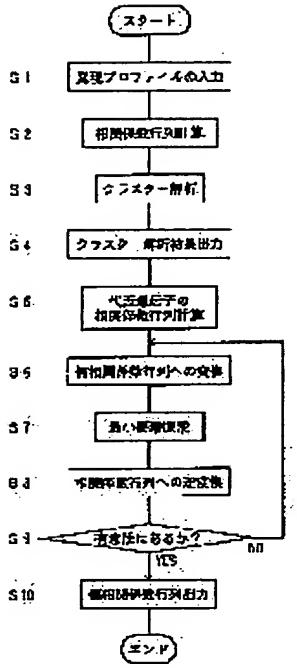
(51)Int.Cl. G06F 17/30
C12M 1/00
C12N 15/09
C12Q 1/68

(21)Application number : 2000-370947
(22)Date of filing : 06.12.2000

(71)Applicant : BIOMOLECULAR ENGINEERING RESEARCH INSTITUTE
(72)Inventor : FUJI HIROYUKI
HORIMOTO KATSUHISA

(54) GRAPHICAL MODELING METHOD AND DEVICE FOR INFERRING GENE NETWORK

(57)Abstract:
PROBLEM TO BE SOLVED: To provide graphical modeling method and device, capable of quickly and precisely inferring a gene network from a manifestation profile.
SOLUTION: The method includes steps in which a gene manifestation profile is inputted, correlation coefficient matrix for genes is calculated, cluster analysis is executed, a cluster analysis result is outputted, and the correlation coefficient matrix of representative genes is calculated. Then transformation into partial correlation coefficient matrix is executed, minimum elements are retrieved, inverse transformation into the correlation coefficient matrix is executed, and significance is determined. When there is no significance, as a result of significance judgment, operation is returned to the transformation into the partial correlation coefficient matrix, but when there is significance, a partial correlation coefficient matrix is outputted, to execute graphical modeling for inferring the gene network.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Japanese Laid-Open Patent Publication
N . 175305/2002 (*Tokukai* 2002-175305)

A. Relevance of the Above-identified Document

The following is a partial English translation of exemplary portions of non-English language information that may be relevant to the issue of patentability of the claims of the present application.

B. Translation of the Relevant Passages of the Document

See also the attached English Abstract.

[INDUSTRIAL FIELD OF THE INVENTION]

The present invention relates to a graphical modeling method for estimating a gene network, and an apparatus used therefor.

[0002]

[PRIOR ART]

With the recent advance in the DNA micro array technique, it is now possible to simultaneously measure expression of thousands of genes under different conditions, which include different stages of cell cycle, different patterns of cell differentiation, diversity of somatic cells, different clinical conditions, and different species. Some of these different conditions can be ordered. For example, different stages of cell cycle or cell differentiation can be ordered on a time scale.

[0003]

As used herein, the "gene expression profile" is used to refer to the expression level of a group of genes measured under different conditions. By evaluating patterns of expression profile, important information can be obtained concerning gene functions and regulation mechanisms. For example, there has

been developed a method of cluster analysis, in which genes on a microarray are clustered based on similarities of expression profiles. As used herein, the "cluster analysis" refers to the process in which groups of genes that show similar expression patterns on a microarray under different measurement conditions are identified and sorted in clusters. For example, the hierarchical clustering (Eisen et al. 1998), the self-organizing mapping (Tomaya et al. 1999), and other clustering methods (Ben-Dor et al. 1999) have been applied to expression profile data.

[0004]

Gene clustering in expression profiling is useful in predicting functions of gene products, or identifying groups of genes that are regulated by the same mechanism.

[0005]

One of the important information contained in the gene expression profile is the expression network of genes. The expression level of a gene is controlled by other genes, either directly or indirectly. Here, such a network of genes is referred to as a "gene network."

[0006]

Estimating a gene network from expression profiles is one of the important goals of functional genomics. The following methods have been addressed for the estimation of a gene network: Boolean network (Somogyi and Shiegoshi, 1996), a differential equation method (Chen et al., 1999; D'haeseleer et al., 1999), and modeling using a combination of these methods (Akutsu et al., 1998).

[0007]

[PROBLEMS TO BE SOLVED BY THE INVENTION]

Conventional methods include estimation using differential equations in combination with gene disruption

experiments, and modeling based on Boolean modeling. However, neither of these methods has been put to actual application.

[0008]

The present invention was made in view of the foregoing problems, and an object of the present invention is to provide a graphical modeling method that can quickly and accurately estimate a gene network from expression profiles. The invention also provides an apparatus for such a graphical modeling method.

[0009]

[MEANS TO SOLVE THE PROBLEMS]

In order to achieve the foregoing objects, the present invention provides (1) a graphical modeling method for estimating a gene network, the method including the steps of: entering a gene expression profile; calculating a gene correlation coefficient matrix; performing a cluster analysis; outputting a result of cluster analysis; performing calculations for a correlation coefficient matrix of representative genes; converting the correlation coefficient matrix into a partial correlation coefficient matrix; searching a minimum element; re-converting the partial correlation coefficient matrix into a correlation coefficient matrix; determining a significance; and returning to step (f) if there is no significance, and outputting the partial correlation coefficient matrix if there is a significance.

[0010]

In order to achieve the foregoing objects, the present invention provides (2) a graphical modeling apparatus for estimating a gene network, the apparatus including: an input section for entering a gene expression profile; a calculation section for calculating a gene correlation coefficient matrix; a cluster analyzing section; a calculation section for calculating a

correlation coefficient matrix of representative genes; a converting section for converting the correlation coefficient matrix into a partial correlation coefficient matrix; a searching section for searching a minimum element; a re-converting section for converting the partial correlation coefficient matrix into a correlation coefficient matrix; a determining section for determining a significance; and an output section for outputting the partial correlation coefficient matrix, the converting section being operated if there is no significance, and the output section outputting the partial correlation coefficient matrix if there is a significance.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－175305
（P2002－175305A）
(43)公開日　平成14年６月21日(2002.6.21)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　17/30	１７０	Ｇ０６Ｆ　17/30	１７０Ｆ	４Ｂ０２４
	２１０		２１０Ｄ	４Ｂ０２９
Ｃ１２Ｍ　1/00		Ｃ１２Ｍ　1/00	Ａ	４Ｂ０６３
Ｃ１２Ｎ　15/09		Ｃ１２Ｑ　1/68	Ｚ	５Ｂ０７５
Ｃ１２Ｑ　1/68		Ｃ１２Ｎ　15/00	Ａ	

審査請求　未請求　請求項の数２　ＯＬ　（全 11 頁）

(21)出願番号　特願2000－370947(P2000－370947)

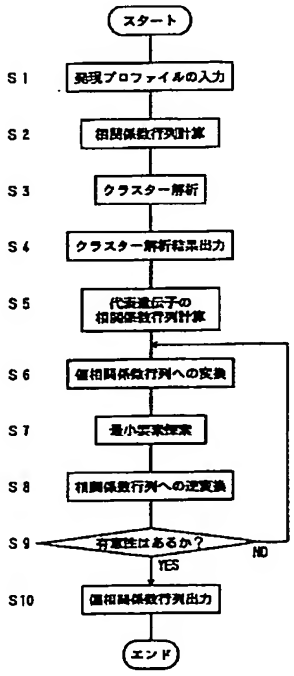
(22)出願日　平成12年12月６日(2000.12.6)

(71)出願人　596013626
株式会社生物分子工学研究所
大阪府吹田市古江台６丁目２番３号
(72)発明者　藤　博幸
大阪府吹田市津雲台３－１　Ａ19－305
(72)発明者　堀本　勝久
佐賀県佐賀市八戸溝３－10－625
(74)代理人　100089635
弁理士　清水　守　（外１名）

Ｆターム(参考)　4B024 AA11 AA20 HA11 HA19
4B029 AA23 BB20
4B063 QA01 QA18 QQ42 QR55 QR84
QS34 QS36 QS39
5B075 ND20 NR12 PR06 UU28

(54)【発明の名称】　遺伝子ネットワークを推測するためのグラフィカルモデリング法及びそのための装置

(57)【要約】
【課題】　発現プロファイルから遺伝子ネットワークを推測する遺伝子ネットワークを迅速、かつ的確に推測することができるグラフィカルモデリング法及びそのための装置を提供する。
【解決手段】　遺伝子の発現プロファイルを入力し、遺伝子の相関係数行列の計算を行い、クラスター解析を行い、クラスター解析結果を出力し、代表遺伝子の相関係数行列の計算を行い、偏相関係数行列への変換を行い、最小要素の探索を行い、相関係数行列への逆変換を行い、有意性の判定を行い、有意性の判定の結果、有意性無しの場合には前記偏相関係数行列への変換へ戻り、有意性有りの場合には偏相関係数行列の出力を行い、遺伝子ネットワークを推測するためのグラフィカルモデリングを行う。



1

【特許請求の範囲】

【請求項１】　遺伝子ネットワークを推測するためのグラフィカルモデリング法において、（ａ）遺伝子の発現プロファイルを入力する過程と、（ｂ）遺伝子の相関係数行列の計算を行う過程と、（ｃ）クラスター解析を行う過程と、（ｄ）クラスター解析結果を出力する過程と、（ｅ）代表遺伝子の相関係数行列の計算を行う過程と、（ｆ）偏相関係数行列への変換を行う過程と、（ｇ）最小要素の探索を行う過程と、（ｈ）相関係数行列への逆変換を行う過程と、（ｉ）有意性の判定を行う過程と、（ｊ）有意性の判定の結果、有意性無しの場合には前記過程（ｆ）へ戻り、有意性有りの場合には偏相関係数行列の出力を行う過程とを施すことを特徴とするグラフィカルモデリング法。

【請求項２】　遺伝子ネットワークを推測するためのグラフィカルモデリング装置において、（ａ）遺伝子の発現プロファイルの入力部と、（ｂ）遺伝子の相関係数行列の計算部と、（ｃ）クラスター解析部と、（ｄ）代表遺伝子の相関係数行列の計算部と、（ｅ）偏相関係数行列への変換部と、（ｆ）最小要素の探索部と、（ｇ）相関係数行列への逆変換部と、（ｈ）有意性の判定部と、（ｉ）偏相関係数行列の出力部とを備え、（ｊ）前記有意性の判定部による有意性の判定の結果、有意性無しの場合には前記変換部（ｅ）へ戻り、有意性有りの場合には前記偏相関係数行列の出力部において出力することを特徴とするグラフィカルモデリング装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、遺伝子ネットワークを推測するためのグラフィカルモデリング法及びそのための装置に関するものである。

【０００２】

【従来の技術】最近のＤＮＡマイクロアレイ技術の進歩により、異なった状況下で数千の遺伝子の発現を同時に測定することが可能になった。ここでいう異なった状況とは細胞周期の各段階、細胞分化の違い、体細胞の多様性、異なった臨床上での条件あるいは異なった種を指す。この中の幾つかの状況には、順番をつけることが可能である。例えば、細胞周期や細胞分化の段階は時間経過に関して順番つけることができる。

【０００３】様々な状況下で測定される一群の遺伝子の発現レベルを、ここでは遺伝子の発現プロファイルと呼ぶ。この発現プロファイルのパターンの評価により遺伝子機能や調節機構に関する重要な知見が得られる。例えば、幾つかのグループは、マイクロアレイ上の遺伝子について、その発現プロファイルの類似性に基づきクラスター分析を行う方法を開発してきている。ここでいうクラスター分析とは、マイクロアレイ上の遺伝子が、異なる測定状況にわたって類似した発現パターンを示す遺伝子群を同定し、クラスターとして分類することをいう。

2

例えば、階層的クラスタリング（Ｅｉｓｅｎ　ｅｔ　ａｌ．１９９８）や自己組織化マッピング（Ｔｏｍａｙａ　ｅｔ　ａｌ．１９９９）や他のクラスタリング方法（Ｂｅｎ－Ｄｏｒ　ｅｔ　ａｌ．１９９９）が発現プロファイルデータに応用されてきている。

【０００４】発現プロファイルによる遺伝子のクラスター化は、機能が未知の遺伝子産物の機能予測や、同じ機構で調節されている遺伝子群を同定するのに役立つ。

【０００５】遺伝子発現プロファイルの中に含まれる重要な情報の一つに、遺伝子間の発現のネットワークがある。一つの遺伝子の発現レベルは、直接的あるいは間接的に他の遺伝子に調節されている。ここではこの様な遺伝子間のネットワークを「遺伝子ネットワーク」と呼ぶ。

【０００６】ある発現プロファイルから遺伝子ネットワークを推測することは機能ゲノム学の重要な課題である。Ｂｏｏｌｅａｎネットワーク（Ｓｏｍｏｇｙｉ　ａｎｄＳｈｉｅｇｏｓｋｉ　１９９６）や微分方程式（Ｃｈｅｎ　ｅｔ　ａｌ．，１９９９；Ｄ’ｈａｅｓｅｌｅｅｒ　ｅｔ　ａｌ．１９９９）やこれらの方法の組み合わせを用いたモデリング（Ａｋｕｔｓｕ　ｅｔ　ａｌ．１９９８）が遺伝子ネットワークの推測のために検討されてきている。

【０００７】

【発明が解決しようとする課題】しかしながら、従来の方法では、遺伝子破壊実験との組合せによる微分方程式による推定や、ブーリアンモデルによるモデル化が行われていたが、実用の段階にはなかった。

【０００８】本発明は、上記状況に鑑みて、発現プロファイルから遺伝子ネットワークを推測する遺伝子ネットワークを迅速、かつ的確に推測することができるグラフィカルモデリング法及びそのための装置を提供することを目的とするものである。

【０００９】

【課題を解決するための手段】本発明は、上記目的を達成するために、

〔１〕遺伝子ネットワークを推測するためのグラフィカルモデリング法において、遺伝子の発現プロファイルを入力する過程と、遺伝子の相関係数行列の計算を行う過程と、クラスター解析を行う過程と、クラスター解析結果を出力する過程と、代表遺伝子の相関係数行列の計算を行う過程と、偏相関係数行列への変換を行う過程と、最小要素の探索を行う過程と、相関係数行列への逆変換を行う過程と、有意性の判定を行う過程と、有意性の判定の結果、有意性無しの場合には前記過程（ｆ）へ戻り、有意性有りの場合には偏相関係数行列の出力を行う過程とを施すことを特徴とする。

【００１０】〔２〕遺伝子ネットワークを推測するためのグラフィカルモデリング装置において、遺伝子の発現プロファイルの入力部と、遺伝子の相関係数行列の計算

部と、クラスター解析部と、代表遺伝子の相関係数行列の計算部と、偏相関係数行列への変換部と、最小要素の探索部と、相関係数行列への逆変換部と、有意性の判定部と、偏相関係数行列の出力部とを備え、前記有意性の判定部の有意性の判定の結果、有意性無しの場合には前記変換部（ｅ）へ戻り、有意性有りの場合には前記偏相関係数行列の出力部において出力することを特徴とする。

【００１１】

【発明の実施の形態】以下、本発明の実施の形態について詳細に説明する。

【００１２】本発明では、特に、発現プロファイルから遺伝子ネットワークを推測する新しい方法を示す。これは「グラフィカルモデリング」と呼ばれている方法を応用したものである。グラフィカルモデリングとは多変量解析のための技術の一つで、グラフによって表現された統計的なモデルを推測あるいは検定するために用いられている。本発明では、この「グラフィカルモデリング」の特徴を活かし、またクラスター分析との組み合わせで新規の遺伝子ネットワーク推定の方法を開発した。この方法では、従来の微分方程式を利用した方法のように遺伝子破壊実験などの組み合わせを必要としないという特徴を持つ。つまり、本発明は、グラフィカルモデリングを発現プロファイルに応用した初めての提案である。

【００１３】遺伝子ネットワークを推測するための方法は、クラスター分析とグラフィカルモデリングの２段階よりなる。まず解析のために、発現プロファイルデータは遺伝子の相関係数行列に変換される。グラフィカルモデリングは共分散選択（Ｄｅｍｐｓｔｅｒ　１９７２）を含んでいる。共分散選択には、相関係数行列の逆行列の計算を要求される。発現プロファイルデータ（Ｅｉｓｅｎ　ｅｔ　ａｌ．１９９８）のクラスター分析に示されるように、多くの遺伝子は類似した発現パターンを示す。発現パターンの類似性のため、相関係数行列のいくつかの列または行の間に疑似的な線形従属の関係が生じる。このために逆行列を数値計算によって求めることが困難になる。

【００１４】この問題を避けるために、まず、クラスター解析を行い、異なるクラスターに属す遺伝子対の間には擬似的な線形従属の関係が生じないようにクラスターの同定を行う。各クラスターから、代表遺伝子を一つずつ選択した。次に選択された遺伝子の発現プロファイルデータから、相関係数行列を求め、その相関係数行列を用いてグラフィカルモデリングを行った。

【００１５】したがって、本発明では、遺伝子間のネットワークではなく、いくつかの遺伝子群の間に形成されるネットワークに関するものである。

【００１６】（発現プロファイルデータ）マイクロアレイ上に、Ｌ個の遺伝子があるものとする。また、発現レベルの計測が行われた状況の数をＮとすると、遺伝子の発現プロファイルは、Ｌ×Ｎのサイズを持つ２次元行列Ｅで表現される。この行列Ｅの要素（ｉ，ｊ）は、ｉ番目の遺伝子のｊ番目の状況における発現レベルを示す。発現プロファイルデータはＳａｃｃｈａｒｉｏｍｙｃｅｓ　ｃｅｒｅｖｉａｅの細胞周期にわたるデータをあるウェブサイト（ｈｔｔｐ：／／ｒａｎａ．ｓｔａｎｆｏｒｄ．ｅｄｕ／ｃｌｕｓｔｅｒｉｎｇ）からとって使用した。

【００１７】（クラスター分析）ここではクラスター分析のため、堀本らの方法を用いた。ただし、クラスター分析はこの方法に制限されるものではない。すなわち、クラスター分析の結果、各クラスターの構成遺伝子の間で擬似的な線形従属性がなくなるような結果が得られるような方法であれば、別の方法によるクラスター分析の結果も利用できる。

【００１８】各クラスターより任意の遺伝子を一個選択し、そのクラスターの代表遺伝子とする。ここでは、Ｌ個の遺伝子に１からＬの番号をふり、各クラスター中で一番小さな番号を有する遺伝子を代表として選択した。

【００１９】（Ｇｒａｐｈｉｃａｌ　Ｇａｕｓｓｉａｎ　ｍｏｄｅｌｉｎｇ）グラフィカルモデリングを理解するためには、条件付きの独立性の概念が重要である。まずこの概念の説明を行う。

【００２０】例として、Ｍ個の遺伝子の集合を考える。このＭ個の遺伝子の発現レベルをある条件下で測定する。ある条件下で計測されたＭ個の遺伝子の発現レベルはＭ次元ベクトル〔ｅｌ（Ｇｅｎｅ　２），ｅｌ（Ｇｅｎｅ３），…，ｅｌ（ＧｅｎｅＭ）〕として表現される。ここでベクトルの要素ｅｌ（遺伝子ｉ）は、遺伝子ｉの発現レベルを示す。ｆはそのＭ次元ベクトルの分布の同時密度関数とする。

【００２１】発現プロファイルに関して遺伝子ＡとＢが相関を示す場合を考えてみる。２つの遺伝子間で相関が生じるための機構として、次の３つの機構が考えられる。

【００２２】第１は、遺伝子間に直接的な相互作用がある場合である。このタイプの相互作用が、遺伝子ネットワークを推定するために必要となるものである。

【００２３】第２は、２つの遺伝子の間に間接的な相互作用がある場合である。言い換えると、遺伝子Ａの産物の調節の情報が、ほかのいくつかの遺伝子の発現を介して、遺伝子Ｂの発現を引き起こすように伝達されている場合である。

【００２４】第３は、共通の遺伝子により調節されているために生じる相関である。

【００２５】遺伝子ネットワークを推測するためには、第一のタイプの相互作用とほかのタイプの相互作用を区別することが重要である。

【００２６】次に、Ｍ次元ベクトルの同時密度関数を考える。この同時密度関数が、一つはｅｌ（遺伝子Ｂ）を

含まず、他方はｅｌ（遺伝子Ａ）を含んでいないような二つの要素に因数分解される時、ｅｌ（遺伝子Ａ）とｅｌ（遺伝子Ｂ）はそれ以外のものが与えられたときに条件付き独立であるという。

【００２７】式：ｆ〔ｅｌ（Ｇｅｎｅｌ），ｅｌ（Ｇｅｎｅ２），…ｅｌ（ＧｅｎｅＡ），…ｅｌ（ＧｅｎｅＢ），…，ｅｌ（ＧｅｎｅＭ）〕＝ｈ〔ｅｌ（ＧｅｎｅＡ），ｔｈｅ　ｒｅｓｔ　〕×ｋ〔ｅｌ（ＧｅｎｅＢ），ｔｈｅ　ｒｅｓｔ〕，

ここで、ｈはｅｌ（遺伝子Ｂ）を含まない関数、ｋはｅｌ（遺伝子Ａ）を含まない関数である。このルールは因数分解基準（Ｄａｗｉｄ，１９７９）とよばれている。条件付き独立性は、２つの発現レベルが直接には関係していないことを意味している。これ以降、記述の煩雑さを避けるため、ｅｌ（遺伝子Ａ）とｅｌ（遺伝子Ｂ）の間に条件付き独立が成立することを、遺伝子Ａと遺伝子Ｂは条件付き独立であると呼ぶことにする。二つの遺伝子が条件付き独立である時には、それらの間に相関があっても直接の相互作用はないことを意味している。　ここで、Ｍ個の遺伝子の間の条件付き独立関係を表現するために、グラフｇ＝（ｖ、ｅ）を考える。ｖは、Ｍ個の遺伝子に対応する頂点の有限集合であり、ｅは２つの頂点間を結ぶ辺の有限集合である。

【００２８】グラフｇは、その発現レベルが条件付き独立であるような遺伝子に対応する頂点間の辺を除去するように因数分解基準を繰り返し適合することによって構築することができる。このとき、ｅは他の要素が与えられた時に条件付き独立でないような発現レベルを有する遺伝子ペア間を結ぶ辺により構成されている。得られたグラフは、Ｍ個の遺伝子の間の遺伝子ネットワークを表現すると見なすことができる。ここでは、先に述べたクラスター分析により、数が減じられた遺伝子の発現プロファイルデータにグラフィカルモデリングを適用することを考える。それらの遺伝子の発現プロファイルデータは、Ｍ×Ｎの２次元実行列で表される。ここで、Ｍはクラスター分析によって減じられた遺伝子の数であり、Ｎは、その発現レベルが計測された条件の数である。いうまでもなく、Ｌ＞Ｍである。モデルを構築するために、Ｍ個の計測値ｅｌ（ＧｅｎｅｌからＭ）は以下の同時分布関数を持つランダムなベクトルであると仮定する。

【００２９】式：ｆθ〔ｅｌ（Ｇｅｎｅｌ），ｅｌ（Ｇｅｎｅ２），…ｅｌ（ＧｅｎｅＡ），…ｅｌ（ＧｅｎｅＢ），…ｅｌ（ＧｅｎｅＭ）〕，

ここで、θは、未知のパラメータである。その時に、先に述べた発現プロファイルを表す２次元行列は、ｆθからの独立なＮ回の観測から得られた標本とみなすことができる。また、パラメータθは、遺伝子の発現レベルの間の条件付き独立性についての情報を含むようにデザインされているものとする。それゆえに、グラフの構造の推定は、パラメータθの推測を通じて行われる。しかし

ながら、ｆやθによる密度関数の記述はあまりにも一般的すぎて、具体的な推定を行うことができない。連続データに関するグラフ構造の推定のためには、ｆやθとして多変量正規分布が仮定される。この時θは、分散行列Σと平均ベクトルμより構成されている。多変量正規分布に基づいた方法は、グラフィカル・ガウシアン・モデリングと呼ばれている。ここでは、Ｍ個の遺伝子の発現レベルは多変量正規分布に従うランダムベクトル〔ｅｌ（Ｇｅｎｅｌ），ｅｌ（Ｇｅｎｅ２），…ｅｌ（ＧｅｎｅＭ）〕であると仮定する。

（独立グラフの推定）多変量正規分布の仮定の下では、因数分解基準はある二変数の間で偏相関係数が０になるということと等価である。すなわち、２つの変数はその偏相関係数が０であるときに限り、条件付き独立となる。偏相関係数を得るためには、共分散行列の逆行列の計算が要求される。

【００３０】（$\Omega=\Sigma^{-1}$）$\Omega=(\omega_{ij})$が与えられたときに、ｅｌ（Ｇｅｎｅｉ）とｅｌ（Ｇｅｎｅｊ）の間の偏相関係数$\rho^{ij,\ the\ rest}$は、Ωの要素から以下のように計算される。$\rho^{ij,\ the\ rest}=-\omega_{ij}/(\omega^{ii}\times\omega^{jj})$として与えられる。この式が示しているように、偏相関係数が０であることは、逆共分散が０であることに対応している。グラフィカル・ガウシアン・モデルは、特定の偏相関係数を０にするようなモデルとして定義される。遺伝子間の条件付き独立の関係を評価することによって、グラフ構造を構築するために、ここでは、ＷｅｒｍｕｔｈとＳｃｈｅｉｄｔらによって開発された、段階的な繰り返しアルゴリズムを用いた。共分散行列や逆共分散行列の代わりに、相関係数行列と偏相関係数行列を用いた。

【００３１】ステップ０：初期化のステップである。完全グラフｇ（０）＝（ｖ，ｅ）を用意する。このグラフの頂点はＭ個の遺伝子に対応しており、全ての頂点は連結されているものとする。このｇ（０）をフルモデルと呼ぶ。発現プロファイルデータに基づいて最初の相関係数行列Ｃ（０）を計算する。ただし、τ＝０とする。

【００３２】ステップ１：相関係数行列Ｃ（τ）から偏相関係数行列Ｐ（τ）を計算する。ここで、τは繰り返しの回数を示す。

【００３３】ステップ２：Ｐ（τ）の全ての要素の中からその絶対値が最小である要素を探して、その要素を０と置き換える。

【００３４】ステップ３：Ｐ（τ）から相関係数行列Ｃ（τ＋１）を再構築する。Ｃ（τ＋１）においては、Ｐ（τ）において０に置き換えられた要素に対応する要素のみが変更される。一方、他のすべての要素は、Ｃ（τ）の要素と同じである。

【００３５】ステップ４：繰り返し計算の打ち切りの判断を行う。そのために以下の３つの数値を計算する。

【００３６】

$$dev1 = N \log\{|C(\tau+1)| / |C(0)|\}$$
$$dev2 = N \log\{|C(\tau+1)| / |C(\tau)|\}$$

|Ｃ（τ＋１）|，|Ｃ（τ）|，|Ｃ（０）|は、各行列の行列式を表す。ｄｅｖ１は自由度＝ｎにおけるχ２分布に従い、ｄｅｖ２は自由度＝１におけるχ２分布に従う。ｎはτ回の反復まで０にセットされた要素の数を示す。Ｎは計測が行われた状況の数を表す。

【００３７】ステップ５：ｄｅｖ１あるいはｄｅｖ２以上の値が生じる確率を、それぞれχ２分布に従い求めた時、ｄｅｖ１に対する確立値が０．５以下か、ｄｅｖ２に対する確立値が０．３以下であれば、Ｃ（τ＋１）に対応するモデルは棄却され、繰り返し計算を終了する。終了条件が満たされなかった時は、Ｐ（τ）において偏相関係数が０にセットされた２つの遺伝子間の辺はｇ（τ）から除去して、ｇ（τ＋１）とし、τをτ＋１に進めて、ステップ１に戻る。

【００３８】上記の方法で最終的に得られたグラフは無向グラフでランダム変数間の条件付き独立関係を表現している。この無向グラフを独立グラフと呼ぶ。

【００３９】以下、本発明の実施例について説明する。

【００４０】図１は本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング装置の構成図である。

【００４１】この図において、１は入力装置、２は発現プロファイル入力部、３は相関係数行列計算部、４はクラスター解析部、５は代表遺伝子の相関係数行列計算部、６はグラフィカル・ガウシアン・モデリング実行部、７は偏相関係数行列への変換部、８は最小要素探索部、９は相関係数行列への逆変換部、１０は有意性判定部、１１は偏相関係数行列出力部、１２は出力装置、１３はＣＰＵ（中央処理装置）、１４は記憶装置である。

【００４２】図２は本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング法のフローチャートである。

【００４３】本発明のグラフィカルモデリング法の概要について図２を参照しながら説明する。

【００４４】①遺伝子の発現プロファイルを入力し（ステップＳ１）、次に、遺伝子の相関係数行列の計算を行い（ステップＳ２）、次に、クラスター解析を行い（ステップＳ３）、次に、クラスター解析結果を出力し（ステップＳ４）、次に、代表遺伝子の相関係数行列の計算を行い（ステップＳ５）、次に、偏相関係数行列への変換を行い（ステップＳ６）、次に、最小要素の探索を行い（ステップＳ７）、次に、相関係数行列への逆変換を行い（ステップＳ８）、次に、有意性の判定を行い（ステップＳ９）、有意性の判定の結果、有意性無しの場合には前記ステップＳ６へ戻り、有意性有りの場合には偏相関係数行列の出力を行う（ステップＳ１０）。

【００４５】以下、具体例を詳細に説明する。

【００４６】（１）クラスター分析

酵母の２，４６７個の遺伝子を、その発現プロファイルに関する相関係数行列の各行あるいは各列が、クラスター間では数値解析の意味において一次独立になるようにクラスターを構成した。

【００４７】その結果、２，４６７個の遺伝子は３４個のクラスターに分類された。

【００４８】（２）グラフィカル・ガウシアン・モデリング（Ｇｒａｐｈｉｘａｌ　Ｇａｕｓｓｉａｎ　Ｍｏｄｅｌｉｎｇ，ＧＧＭ）

上記の３４クラスターの代表遺伝子の発現プロファイルの相関係数行列を用いてＧＧＭを実行した。

【００４９】ＧＧＭの繰り返し計算の過程における、ｄｅｖ１とｄｅｖ２のｐ－ｖａｌｕｅの変化を繰り返しのステップに対しプロットしたものを、図３に示す。すなわち、図３では、ｄｅｖ１とｄｅｖ２の確立値の繰り返しのステップ番号に対してのプロットｄｅｖ１とｄｅｖ２の確立値は、ＧＧＭの繰り返し計算の停止の判定のために使用されており、それを繰り返しのステップ数の関数として、プロットしている。●はｄｅｖ１の確立値、○はｄｅｖ２の確立値を表す。

【００５０】この図から明らかなように、ｄｅｖ１のｐ－ｖａｌｕｅは、図中ａで示すように、全過程において９９．９９％より大きく、繰り返しの打ち切り評価には利用されなかった。一方、ｄｅｖ２のｐ－ｖａｌｕｅは、図中ｂで示すように、繰り返しステップが進むにつれて減少してゆき、１３７ステップ目で偏相関係数行列（ＰＣＣＭ）の要素（９，２８）の評価の際、ｐ－ｖａｌｕｅが０．２８４となり設定した基準値０．３より小さくなったため繰り返しは打ち切られた。したがって、ＰＣＣＭ中１３６個の要素が０．０に置換されたことになる。以上のように、ＧＧＭの繰り返し計算の打ち切りには、ｄｅｖ２の方が効率的であった。

【００５１】ＧＧＭによって得られた３４個の代表遺伝子の最終的なＰＣＣＭを図４に示す。すなわち、図４には、ＧＧＭによって得られた偏相関係数の行列ＧＧＭの繰り返しの過程で０．０に置換された要素は網がけで示されている。

【００５２】要素（７，８）がＰＣＣＭ中最小で、－０．７５、要素（４，２４）が最大で０．５９であった。５６１個のＰＣＣＭの要素の内、１３６個が０に置換された（約２４％）。言い換えると、１３６個の辺が独立グラフから削除されたことになる。独立グラフには孤立した節はなく、また全ての他の節に辺を有する節もなかった。節の内、最多の辺を有するものでは、３０個、最少の辺を有するものでは１８個の辺を有していた。独立グラフで全ての辺を表示するのは煩雑になるため、ここでは偏相関係数の絶対値が０．５より大きなもののみ表示したものを図５に示す。すなわち、図５に

は、３４個のクラスターの独立グラフが示されており、絶対値が０．５より大きな偏相関係数に対応する辺のみが示されている。

【００５３】以下、さらに議論を深めることにする。

【００５４】（１）代表遺伝子のＧＧＭのための妥当性２，４６７個の遺伝子の発現プロファイルデータのサンプル数がさらに増加すれば、より小さなメンバーから構成されたクラスターへの分割が可能になるかもしれない。しかし、ここでは、得られた３４クラスターの各々に含まれる遺伝子は、その発現の挙動が類似している、すなわち類似した制御を受けていると仮定し、３４個の遺伝子のＰＣＣＭ（図４）と独立グラフ（図５）は、３４個のクラスターの間のＰＣＣＭと独立グラフを表しているものとして議論をすすめる。

【００５５】その前に、代表遺伝子の選択の妥当性を述べておく。代表遺伝子を選択する様々のものが考えられるが、そのような方法の違いが、ＧＧＭの結果に影響を与えないか否かについて検討する。

【００５６】まず、先に選択した代表遺伝子以外の３４クラスターのメンバーをランダムに選択して、１００個の３４個の代表遺伝子のセットを用意した。そして、その各々と、先に選択した代表遺伝子について、その発現プロファイルの相関係数行列の行列式の比（比が１より小さい場合はその逆数）の対数に５６１を乗じたものは自由度５６１のカイ自乗分布に従うと予想されるので、そのｐ－ｖａｌｕｅを計算してみた。有意水準を０．０５とすると、１００個中７個のセットのみがオリジナルの代表遺伝子セットとは有意に異なるものとして棄却された。また、ＧＧＭでｄｅｖ１、ｄｅｖ２による繰り返しステップの打ち切りに使用している有意水準、０．３と０．５を使用しても、１００セットの内９セットと１１セットがそれぞれオリジナルの代表遺伝子とは有意に異なるものとして棄却された。この結果は、必ずしも異なる代表遺伝子セットを用いたＧＧＭの結果が一致することを保証はしないものの、一組の代表遺伝子によってＧＧＭを行った結果で十分であることを強く示唆している。

【００５７】（２）独立グラフによる辺の解釈

上記したように、ここは、ＰＣＣＭの非０要素あるいはそれに対応した独立グラフの辺は、クラスター間の直接的な相互作用を表しているものと仮定している。このセクションでは、辺の解釈について吟味する。一般的にＰＣＣＭの非０要素は変数のペアの間の直接的な相互作用を示唆している。しかし、それはどちらの変数が原因でどちらが結果であるかは示すことができない。それ故に、独立グラフ中の辺は無向である。因果関係に対する先験的な知識無しには辺を矢印に置き換えることはできない。図４に示されているＰＣＣＭの非０要素はクラスターのペアの間の直接的な相互作用を表すものと考えられている。そのときに、クラスター間の直接的な相互作用がなにを意味するかについて説明する。

【００５８】あるクラスターに含まれているすべての遺伝子が、辺で結ばれている他のクラスター内のすべての遺伝子の発現に影響するということは現実的であるとは思えない。むしろ、あるクラスターの遺伝子の部分集合が他のクラスターの遺伝子の部分集合の発現に影響しているということの方が考えやすい。更に、クラスター間の辺はループを構成しているかもしれない。例として、１つの辺で結ばれたクラスターＡ，Ｂで考えてみよう。クラスターＡの遺伝子の部分集合にクラスターＢの遺伝子の部分集合が発現に影響を及ぼしており、一方、クラスターＢの他の遺伝子の部分集合がクラスターＡの部分集合に影響を与えていることが考えられている。このように、クラスター間の辺が遺伝子間に及ぼす影響は、直接的な相互作用よりさらに複雑な意味を持っているかもしれない。加えて、ここで検討された発現プロファイルデータには、機能について未知の多くの遺伝子が含まれていないということを心に留めておかなければならないだろう。

【００５９】アイゼンらの仕事（１９９８）においては、その機能が同定されている酵母の２４６７個の遺伝子の発現レベルが計測されている。しかし、ルビンら（２０００）によれば、酵母のゲノム中には６２４１個の遺伝子が蛋白をコードされるものとして存在することが予想されている。このように、データが欠けていることによって、いくつかの辺は直接的な相互作用ではなく、機能未同定の遺伝子の発現を通じて、間接的に相互作用している遺伝子のために生じたという可能性を除去することができない。

【００６０】（３）推定された遺伝子ネットワークの特徴

独立グラフは孤立点を含まなかった。また、どの節も他のすべての節に対して、辺を持つものはなかった。遺伝子ネットワークについての重要な問題の１つは、遺伝子ネットワークはいくつかの独立した部分ネットワークに分割できるかどうかということである。この問題は、独立グラフが他の部分グラフの節に対して辺を持たないようないくつかのサブグラフに分割できるかどうかということで、容易に確認することができる。得られたＰＣＣＭ（図４）の検討から、そのような分割はできないということが示唆された。すなわち、遺伝子ネットワークは単一のシステムとして働いている。推定された遺伝子ネットワークの調節機構を特徴づけるため、ＰＣＣＭをヒストグラムの形に要約して示したものが図６である。すなわち、図６では、偏相関係数の頻度分布が示されている。横軸は偏相関係数を表している。頻度の計算のため、偏相関係数は、０．１の幅を持った離散データにまるめられた。縦軸は、幅内の係数の頻度を表している。しかしながら、偏相関係数の０．０の頻度は横軸上の０．０に対してプロットされている。

【００６１】この図６のヒストグラムでは、偏相関係数の頻度分布が示されている。ＰＣＣＭの５６１個の要素中、２３９個が正の値を持っており、一方、１８６個は負の値を持っていた。そして、その数はそれぞれ４３％、３３％に対応していた。さらに、正の偏相関係数の分布は、負の相関係数の分布に対して、対称ではなかった。正の偏相関係数の分布は、０．２から０．３までの範囲にピークを持っていた。さらに、０．０から０．１の範囲の相関係数は存在しなかった。これに対して、負の偏相関係数の分布のピークは、－０．２から－０．３にあった。分布の端は、正の相関係数の場合に比べ、両側にのびていた。この観察は、正の調節の数の方が負の調節の数よりも多いことを示唆している。

【００６２】ｍＲＮＡの転写に関与している遺伝子は、遺伝子の発現の調節に関わっている可能性が最も高い。それ故にあるクラスター中の転写関連遺伝子の集合は、辺によって結ばれている他のクラスターの遺伝子の発現の調節において中心的な役割を演じているかもしれない。３４のクラスター中の、３３個はｍＲＮＡの転写に関与している可能性を持っていた。しかしながら、その量はクラスターごとに異なっていた。この観察はｍＲＮＡの転写が並列分散型の調節を受けていることということを示唆している。ＰＣＣＭ（図４）や独立グラフ（図５）は、ある程度まではｍＲＮＡの発現に関与している

遺伝子による発現の調節を表しているのかもしれない。一方、クラスター３２は、翻訳関連遺伝子の４０％以上と、ｒＲＮＡの転写関連遺伝子の約５０％を含んでいた。クラスター３２に結ばれている辺の数は２８であり、５つの辺がＧＧＭの過程で除かれたことになる。他のクラスターにおいては、翻訳関連遺伝子の含量は低かった。したがって、翻訳はむしろ、集中的な調節系を採用しているのかもしれない。

【００６３】ここでは、細胞周期調節遺伝子に注目して、本発明のモデルネットワークを簡略化してみた。シュベルマンらは、２９４個の細胞周期調節遺伝子を酵母中で同定していた。この研究に使われた２４６７個の遺伝子のうち、２３２個の遺伝子がシュベルマンらによって分類されている細胞周期調節遺伝子と対応していた。細胞周期のステージはＧ１，Ｓ，Ｇ２，Ｍ，Ｍ／Ｇ１の５つがあり、その各々が周期に特異的に発現している細胞周期庁因子を有している。２３２個の遺伝子のうち、９３，３２，２８，４８，３１個の遺伝子がそれぞれＧ１，Ｓ，Ｇ２，Ｍ，Ｍ／Ｇ１に関連していた。ここで、３４個のクラスターの中におけるそれらの遺伝子の分布を調べてみたものが表１である。

【００６４】

【表１】

クラスター	Ｇ１	Ｓ	Ｇ２	Ｍ	Ｍ／Ｇ１	合計
１	０	０	０	１	０	１
２	１	５	１	１	１	９
３	０	０	０	０	０	０
４	１	１	１	１	０	４
５	０	０	１	０	０	１
６	０	０	０	０	０	０
７	１５	４	１０	４	２	３５
８	１	１	１	０	０	３
９	８	０	１	５	１	１５
１０	０	０	０	１	２	３
１１	１２	９	０	１	２	２４
１２	３６	６	３	２	０	４７
１３	１	０	０	１	０	２
１４	０	１	０	４	１	６
１５	２	０	１	１	０	４
１６	０	１	０	０	１	２
１７	０	０	０	０	０	０
１８	０	０	０	０	２	２
１９	０	０	０	０	１	１
２０	１	１	０	１	３	６
２１	１	０	０	０	０	１
２２	１	０	０	１	１	３
２３	１	０	２	２	４	９
２４	１	０	２	２	１	６
２５	１	０	１	９	０	１１
２６	１	１	０	１	２	５
２７	１	０	０	２	１	４
２８	３	０	０	０	１	４
２９	０	１	０	０	０	１
３０	１	１	０	４	２	８
３１	２	０	０	１	３	６
３２	１	０	２	２	０	５
３３	０	０	０	０	０	０
３４	１	０	２	１	０	４
	９３	３２	２８	４８	３１	２３２

【００６５】３４個のクラスター中、４個は何の細胞周期調節遺伝子も含んでいなかった。クラスター１２は、Ｇ１に関連している細胞周期関連遺伝子を最も多く含んでいた。同様に、クラスター７と２５は、それぞれＧ２とＭ２に関与している細胞周期関連遺伝子を最も多く含んでいた。ＳとＭ／Ｇ４に関連した細胞周期関連遺伝子は、いくつかのクラスターに分散して存在しているようにみえるが、比較的クラスター１１と１２に集中していた。ここではクラスター７，１１，１２，２３，２５に関連したＰＣＣＭの要素を集めてみた。それを図７に示す。すなわち、図７には、クラスター７，１１，１２，２３，２５の間の偏相関係数が示されている。

【００６６】図７に示すように、９つの要素のうち、６個は０だった。また、２つの要素（１１，２５）と（１２，２５）は、０．１１という小さな値を持っていた。クラスター１１と１２のペアは０．３７、クラスター７と２３のペアに対応する要素のみが－０．３２という大きな値を示していた。この観察は、細胞周期の異なるステージに関与している細胞周期調節遺伝子が基本的にはお互いに条件付き独立であるか、非常に弱い関係しか持たないということを示唆している。

【００６７】なお、本発明は上記実施例に限定されるものではなく、本発明の趣旨に基づいて種々の変形が可能であり、これらを本発明の範囲から排除するものではない。

【００６８】

【発明の効果】上記したように、本発明によれば、発現プロファイルから遺伝子ネットワークを推測するグラフィカルモデリングにより遺伝子ネットワークを迅速、かつ的確に推測することができる。

【図面の簡単な説明】

【図１】本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング装置の構成図である。

【図２】本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング法のフローチャートである。

【図３】ＧＧＭの繰り返し計算の過程における、ｄｅｖ

15

１とｄｅｖ２のｐ－ｖａｌｕｅの変化を繰り返しのステップに対しプロットした図である。

【図４】ＧＧＭによって得られた３４個の代表遺伝子の最終的なＰＣＣＭを示す図である。

【図５】３４個のクラスターの独立クラスター。偏相関係数の絶対値が０．５より大きなもののみ表示した図である。０．５～０．５５が異なる長さのセグメントよりなる破線。０．５５～０．６が同じ長さのセグメントよりなる破線。０．６～０．７は細い実線。０．７～０．８は太い実線で示されている。

【図６】推定された遺伝子ネットワークの調節機構を特徴づけるため、ＰＣＣＭをヒストグラムの形に要約して示した図である。横軸は偏相関係数を表している。頻度の計算のため、偏相関係数は、０．１の幅を持った離散データにまるめられた。縦軸は、幅内の係数の頻度を表している。しかしながら、偏相関係数の０．０の頻度は横軸上の０．０に対してプロットされている。

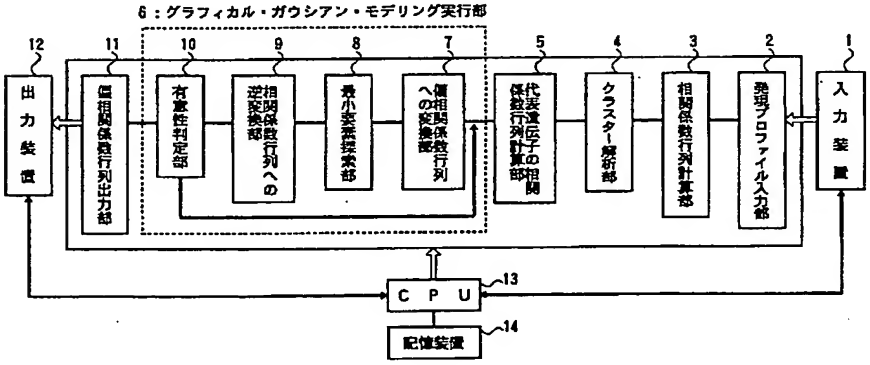
16

＊【図７】クラスター７，１１，１２，２３，２５に関連したＰＣＣＭの要素を集めた図である。

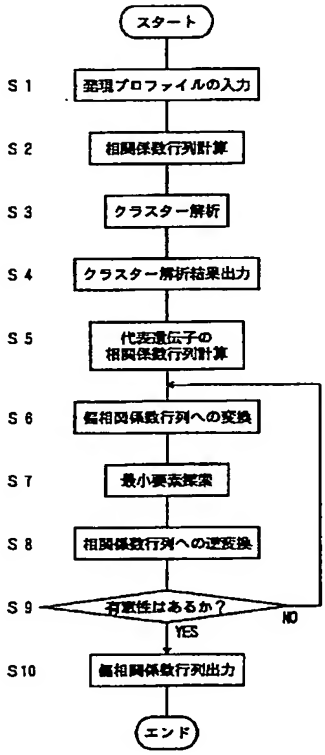
【符号の説明】

１　入力装置
２　発現プロファイル入力部
３　相関係数行列計算部
４　クラスター解析部
５　代表遺伝子の相関係数行列計算部
６　グラフィカル・ガウシアン・モデリング実行部
７　偏相関係数行列への変換部
８　最小要素探索部
９　相関係数行列への逆変換部
１０　有意性判定部
１１　偏相関係数行列出力部
１２　出力装置
１３　ＣＰＵ（中央処理装置）
１４　記憶装置

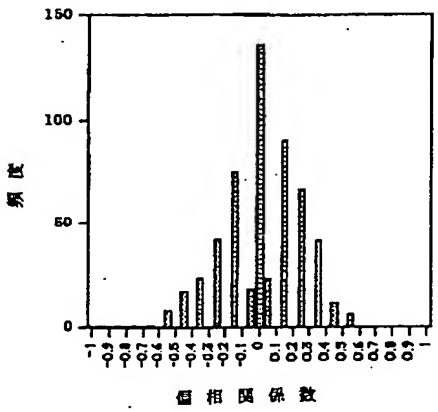
【図１】



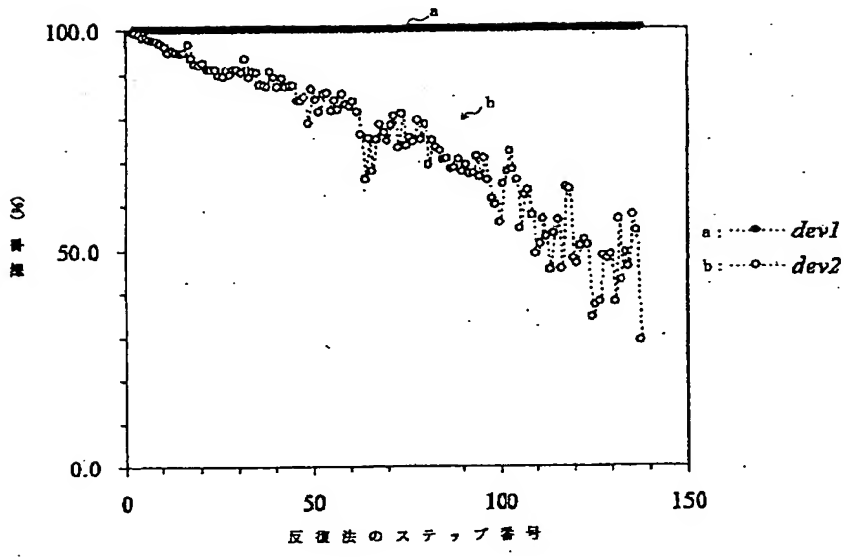
【図２】



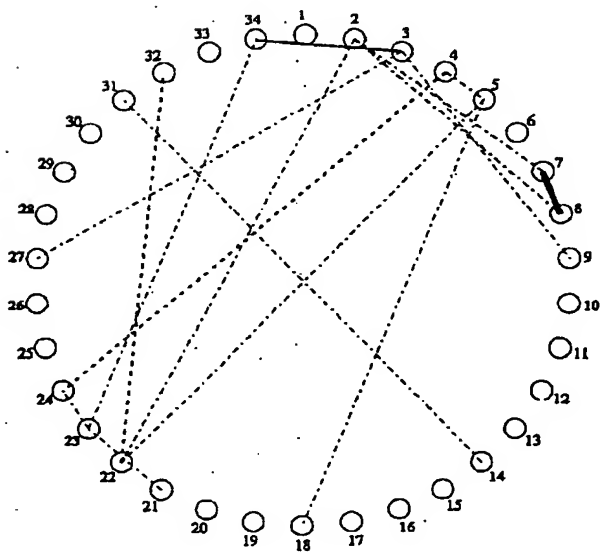
【図６】



【図３】



【図５】



【図７】

G1	12	－				
S	11	0.37	－			
G2	7	0.00	0.00	－		
M	25	0.11	0.11	0.03	－	
M/G1	23	0.00	0.03	－0.32	0.00	－
		12	11	7	25	23

【図４】